INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust Vowel Landmark Detection using Epoch-based Features

Sri Harsha Dumpala, Bhanu Teja Nellore, Raghu Ram Nevali,
Suryakanth V. Gangashetty and B. Yegnanarayana

International Institute of Information Technology, Hyderabad, India
{sriharsha.dumpala, bhanuteja.nellore, raghuram.nevali}@research.iiit.ac.in,
svg@iiit.ac.in, yegna@iiit.ac.in

Abstract

Automatic detection of vowel landmarks is useful in many applications such as automatic speech recognition (ASR), audio search, syllabification of speech and expressive speech processing. In this paper, acoustic features extracted around epochs are proposed for detection of vowel landmarks in continuous speech. These features are based on zero frequency filtering (ZFF) and single frequency filtering (SFF) analyses of speech. Excitation source based features are extracted using ZFF method and vocal tract system based features are extracted using SFF method. Based on these features, a rule-based algorithm is developed for vowel landmark detection (VLD). Performance of the proposed VLD algorithm is studied on three different databases namely, TIMIT (read), NTIMIT (channel degraded) and Switchboard corpus (conversational speech). Results show that the proposed algorithm performs equally well compared to state-of-the-art techniques on TIMIT and better on NTIMIT and Switchboard corpora. Proposed algorithm also displays consistent performance on TIMIT and NTIMIT datasets for different levels of noise degradations.

Index Terms: Vowel landmarks, epochs, excitation source, zero frequency filtering, single frequency filtering.

1. Introduction

Vowels are produced with a relatively open vocal tract compared to the adjacent consonants, resulting in higher spectral amplitude in the first and second formant frequency ranges of vowels than the corresponding spectral amplitude in adjacent consonants [1]. This causes a maximum in the first formant frequency spectral amplitude during the production of vowels, which serves as an evidence for vowel landmarks [1, 2]. The problem of vowel landmark detection (VLD) is to detect the location of the vowel landmarks in the speech signal. Acoustically, vowels (especially, regions around vowel landmarks) have higher intensity than consonants [1 - 3]. This makes VLD useful in building robust automatic speech recognition (ASR) systems and voice activity detection systems. Also, tasks such as syllabification of speech, speech rate estimation and language identification are based on VLD [8 - 14]. These tasks require to deal with both, read and conversational speech collected in different environments. Hence, the motivation for robust VLD.

Approaches for VLD were proposed in different contexts [2, 6 - 14]. VLD plays a crucial role in developing landmark-based ASR systems [2, 6, 7]. VLD method based on Mermelstein's convex hull algorithm was proposed in [2], where energy peaks in the first formant range (0 - 900 Hz) were considered as vowel landmarks. In [6], vowel landmarks were obtained by combining peaks detected in the energy contours of different frequency bands. Local peaks detected in the output of the support vector machine (SVM) trained with mel frequency cepstral coefficients (MFCCs) for the task of vowel classification were considered as vowel landmarks [7].

In literature, vowel landmarks were also considered as syllable nuclei [2, 9 - 13]. Hence, VLD based methods were proposed for syllable detection and estimation of speaking rate [9 - 14]. VLD performed by detecting peaks in energy contour, periodicity measure and instantaneous speech rhythm was considered for syllable detection [9 - 10]. Peaks in the linear prediction residual combined with the formant peaks, obtained using group-delay spectrum, were used to obtain vowel landmarks [11]. Syllabification of conversational speech was performed by detecting vowel landmarks using bi-directional long-short-term memory (BLSTM) neural networks [12]. The BLSTM neural network was trained with sub-band modulation spectrum values and perceptual linear prediction (PLP) coefficients. VLD based on smoothed modified loudness contour was used for speaking rate estimation [13]. Spectral sub-band correlation methods, extended by including temporal correlation and by using prominent spectral bands, were considered for VLD to estimate speaking rate [14].

Vowel landmarks can be characterized by both excitation source and vocal tract system based features, but most of the existing methods for VLD are based on spectral features. The main objective of this study is to develop a robust VLD algorithm using acoustic features which represent both excitation source and vocal tract system characteristics of vowels. In this work, all the proposed features are extracted around epoch locations. Epochs are the instants of significant excitation of the vocal tract system which correspond to the glottal closure instants of the glottal cycle [15]. Generally, regions around epochs have high signal-to-noise ratio (SNR) values, and are relatively more robust to external degradations than other regions of speech [15, 16]. Hence, features extracted around epochs are more reliable in presence of noise [17].

The paper is organized as follows. Section 2 describes the different datasets considered for analysis. Methods used for extraction of epochs and the proposed acoustic features are explained in Section 3. Acoustic features along with the proposed approach for VLD are described in Section 4. Results are discussed in Section 5. Final section gives the summary and conclusions of this work.

2. Description of databases

Three different databases namely, TIMIT [18], NTIMIT [19] and ICSI switchboard corpus [20] are considered to test the proposed approach for VLD.

The proposed approach is developed and evaluated by con-

http://dx.doi.org/10.21437/Interspeech.2016-1074

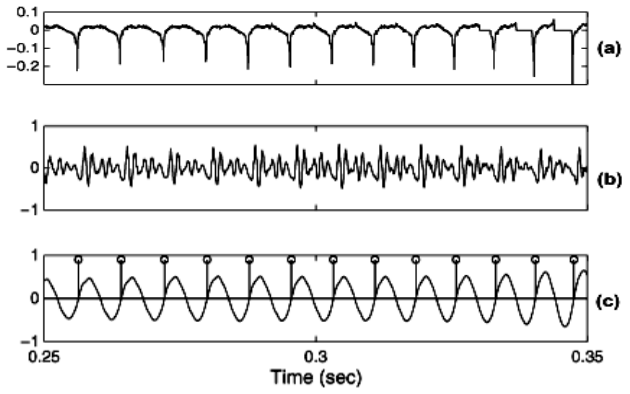

Figure 1: Illustration of the ZFF method for epoch extraction. (a) Differenced EGG (dEGG) signal of a vowel segment, (b) speech waveform of the vowel segment, (c) ZFF signal obtained from the speech signal with hypothesized epoch locations at positive zero crossings.

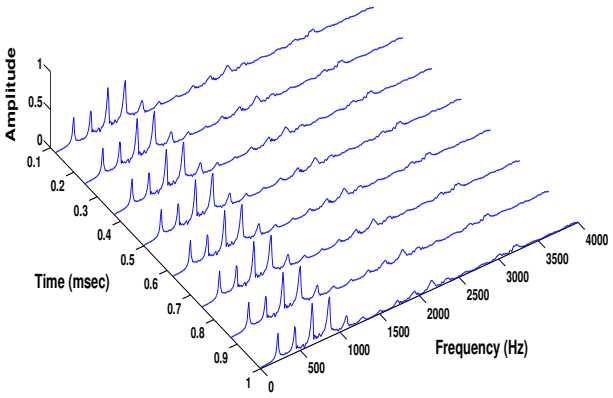

Figure 2: The envelopes obtained using SFF method at every 10 Hz in the frequency range of 0 Hz - 4000 Hz for a vowel segment.

sidering a subset of the standard TIMIT database. To set the parameters for the proposed algorithm, 100 utterances from TIMIT training set, spoken by 25 speakers (15 male and 10 female) are used. 1000 utterances, spoken by 168 speakers (112 male and 56 female) from the TIMIT test set are used to evaluate the proposed VLD algorithm.

To study the performance of the proposed algorithm against channel degradations, the NTIMIT database is used. 1000 utterances in the NTIMIT test set are considered to evaluate the VLD algorithm.

The ICSI switchboard (STP) corpus is considered to evaluate the performance of the proposed approach on conversational speech. This corpus consists of large number of telephone conversation recordings and is considered as a fair representative of spontaneous speech. A total of 500 utterances each of 1 to 10 seconds duration are considered to evaluate the proposed VLD algorithm.

In order to study the robustness of the proposed approach, white noise from NOISEX database [21] is added to TIMIT and NTIMIT datasets at various noise levels.

3. Epoch and acoustic feature extraction methods

3.1. Zero frequency filtering (ZFF) method

Features representing the glottal source of excitation are extracted directly from the speech signal using zero frequency filtering (ZFF) method [15]. The ZFF method is used to estimate the location of epochs by passing the speech signal through a cascade of two zero frequency resonators (ZFR). The trend in the output of ZFR is removed by local mean subtraction using a window length of about 1.5 times the average pitch period (computed using autocorrelation function). This trend removed output is called zero frequency filtered (ZFF) signal. The instants of negative to positive zero crossings of the ZFF signal are called epochs. This method of epoch extraction was shown to be robust against different types of degradations even at low SNRs [15]. An illustration of ZFF method on a vowel segment is shown in Fig. 1. ZFF signal (obtained for the vowel segment in Fig. 1(b)) along with the epoch locations is shown in Fig. 1(c). Fig. 1(a) shows the differenced EGG (dEGG) signal which is used as a reference for the epoch locations.

3.2. Single frequency filtering (SFF) method

Features capturing vocal tract system characteristics are extracted using recently proposed single frequency filtering (SFF) technique [23]. Using SFF method, the amplitude envelope of the speech signal can be obtained with high spectral and temporal resolution at each frequency. In this method, the envelope is obtained at any frequency by frequency shifting the speech signal and filtering the resulting signal using a single-pole filter. The root of the single pole filter is located on the unit circle at the highest frequency, i.e., at $f_s/2$, where f_s is the sampling frequency. Brief description of the steps to extract the envelope of the signal at any desired frequency f_k is as follows [23]:

1. Difference the input speech signal $s[n]$.
$$x[n] = s[n] - s[n-1]. \quad (1)$$
2. Multiply the speech signal $x[n]$ with a complex sinusoid $e^{j\bar{w}_k n}$, where $\bar{w}_k = \pi - w_k = \pi - \frac{2\pi f_k}{f_s}$, to shift the frequency spectrum $X(w)$ of the signal $x[n]$.
The resulting frequency shifted signal
$$x_k[n] = x[n]e^{j\bar{w}_k n}, \quad (2)$$
is passed through a single-pole filter whose transfer function is *H(z)*, where
$$H(z) = \frac{1}{1 + rz^{-1}}. \quad (3)$$
This filter has a pole on the real axis at a distance of r from the origin. Hence the root at $z = -r$ in the z-plane is set such that it corresponds to $f_s/2$.
3. The output $y_k[n]$ of the filter is given by
$$y_k[n] = -ry_k[n-1] + x_k[n]. \quad (4)$$
4. The envelope of the signal $y_k[n]$ is given by
$$v_k[n] = \sqrt{y_{kr}[n]^2 + y_{ki}[n]^2}, \quad (5)$$
where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$, respectively.

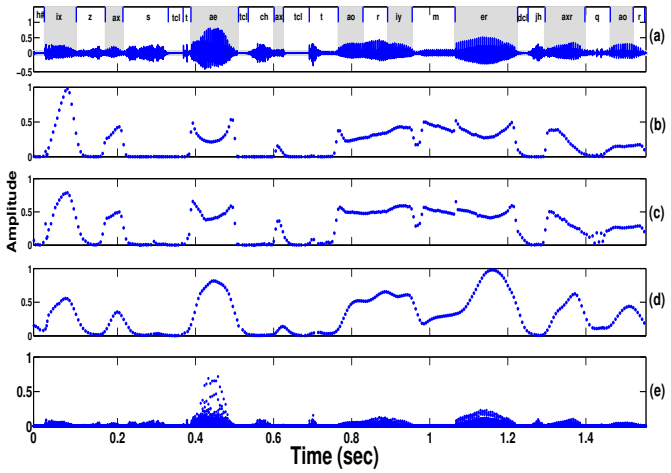

Figure 3: Features considered for Vowel landmark detection (a) Speech waveform for the TIMIT utterance "Is a statutory merger". Shaded regions in manual phone labels represent vowels. (b) α, (c) β, (d) γ values around epochs. (e) S_v values at every sample of the speech signal.

The value of the single-pole filter i.e., r is chosen to be 0.99, but not exactly 1, to ensure stability in the filter output. In this study, the envelope is computed at every 10 Hz in the range of 0 Hz to $f_s/2$ i.e., 4000 Hz as a function of time. The envelopes obtained at every 10 Hz for a vowel segment at every sampling instant are shown in Fig. 2.

3.3. Proposed features for VLD

Excitation source based features extracted from the ZFF signal for VLD are given below.

ZFF signal energy (α): The α value is computed as the energy of the ZFF signal within a window of 2 msec around each epoch (1 msec on each side of epoch). The absence of any acoustic discontinuity during the production of vowels allows free vibration of the vocal folds [1]. This results in higher energy concentration around epochs in vowels, which is reflected by the higher values of α in vowel regions compared to adjacent consonants as shown in Fig. 3(b).

Strength of excitation (β): The β values correspond to the rate of glottal closure and are proportional to the slope of the ZFF signal around epochs [22]. During the production of vowels, there is no significant pressure drop in the subglottal airway resulting in stronger excitation of the vocal tract system [1]. This is represented by the higher values of β in vowel regions compared to adjacent consonants as shown in Fig. 3(c).

Following features are extracted from the envelopes $v_k[n]$ to detect vowel landmarks.

Dominant resonance strength (γ): γ represents the maximum amplitude across all the envelopes at a given epoch location and is computed as

$$\gamma[t_i] = \max_{1 \le k \le N_{k1}} v_k[t_i], \tag{6}$$

where N_{k1} refers to the number of envelopes in the required frequency range and t_i is the i^{th} epoch location. In this paper, frequency range of 0 Hz to 900 Hz is considered to compute γ values as the spectral energy is higher for vowels in this range [2]. Hence, vowel regions exhibit higher γ values compared to adjacent consonants as shown in Fig. 3(d). In this study, local peaks detected in the γ contour are considered as the initially hypothesized vowel landmarks.

Spectral variance (S_v): S_v is the variance of the envelopes across different frequencies, which is computed as

$$S_v[n] = \frac{\sum_{k=1}^{N_k} (v_k[n] - \mu[n])^2}{N_k}, \tag{7}$$

where $\mu[n]$ is the mean of all frequency envelopes at the time instant n and N_k is the number of frequencies considered (here N_k = 400, as frequencies at every 10 Hz in the range of 0 Hz to 4000 Hz are considered). Generally, vowels have larger dynamic range in the frequency domain compared to the consonants [24]. This is represented by the high S_v values in the vowel regions compared to consonants as shown in Fig. 3(e), where S_v values are obtained at every sampling instant.

4. Algorithm for vowel landmark detection

A rule-based algorithm is developed to detect vowel landmarks in continuous speech. Features discussed in this section are employed to develop the algorithm for VLD. All features values are normalized between 0 and 1 (amplitude normalization). The thresholds laid on the features are selected using histogram based analysis performed on 100 speech utterances from TIMIT training set. Steps in the proposed VLD algorithm are as follows.

1. Extract epoch locations from the speech signal.
2. Compute γ at every epoch location.
3. The first set of evidences for vowel landmarks are obtained by picking the peaks in the γ contour as shown in Fig. 4(b). Except few, most peaks represent vowel landmarks.
4. A threshold of 0.015 is used for the peak values. Only peaks with γ values above this threshold are retained.
5. Compute α and β at every epoch location.
6. A threshold of 0.02 and 0.05 is used for α and β values, respectively. Select only epochs with α and β values above these thresholds as shown in Fig. 4(c) for further analysis.
7. Among the peaks retained after step 4, consider only those peaks belonging to epochs obtained in step 6 as shown in Fig. 3(d).
8. A threshold of 5 X 10^{-4} is used for S_v values obtained at every sampling instant. Regions with S_v above the threshold for more than 10 msec are considered potential for VLD (see Fig. 4(e)). Only peaks occurring in these regions are retained as shown in Fig. 4(f).
9. If the amplitude of the valley (minimum value) between two adjacent peaks is greater than 50 percent of the maximum value of any of the two peaks, then the peak with lower amplitude is eliminated. This ensures that more than one peak is not picked in a single vowel segment.
10. Based on the assumption that the minimum distance between two spectral peaks is not less than 20 msec, if any two peaks are separated by less than 20 msec duration, then the peak with the lower amplitude is eliminated.

Final set of peaks, considered as vowel landmarks, obtained by the proposed algorithm are shown in Fig. 4(g). The arrows

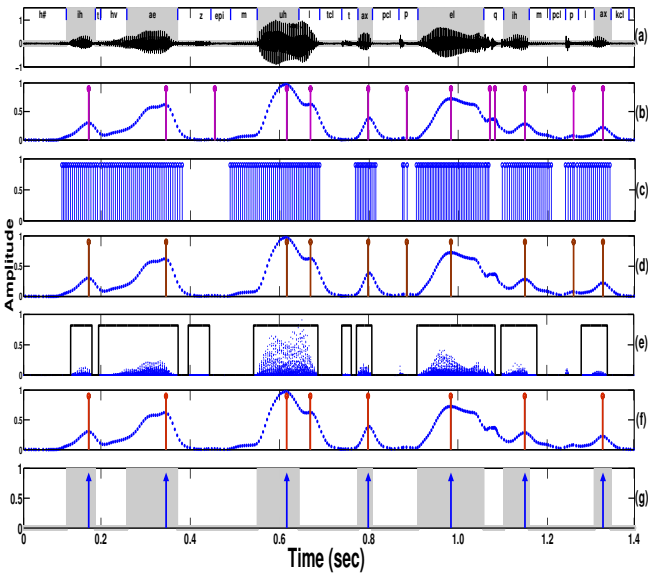

Figure 4: Algorithm for Vowel landmark detection (a) Speech waveform for the TIMIT utterance "it has multiple implica...", (b) vertical lines represent the peak locations in the γ contour, (c) epoch locations obtained by using thresholds on α and β values, (d) peaks retained after validating with α and β values, (e) decision obtained by using threshold on S_v, (f) peaks retained after validating with S_v values, (g) final set of peak (vowel landmark) locations are represented by the arrows. The shaded portions in (a) and (g) correspond to the vowel regions as obtained from manual boundaries.

correspond to the detected vowel landmark locations and the shaded portions represent the manual vowel boundaries which serve as ground truth.

5. Results and discussion

Performance of the proposed VLD algorithm is evaluated on the datasets described in Section 2. Phonetic transcriptions provided for the datasets are used as ground truth for vowel phones. If a detected landmark lies in the vowel region, it is considered as hit. Performance is evaluated in terms of recall, precision and F-measure. Recall is defined as the ratio of the number of hits to the number of vowel segments in the ground truth. Precision is defined as the ratio of the number of hits to the number of landmarks detected. F-measure is the harmonic mean of the recall and precision.

Table 1: *Performance evaluation on TIMIT dataset.*

Measure	BLSTM	PSF	Proposed
Recall	92.22	92.67	91.71
Precision	95.82	91.06	94.77
F-measure	93.98	91.86	93.21

Table 2: *Performance evaluation on NTIMIT dataset.*

Measure	BLSTM	PSF	Proposed
Recall	87.34	91.46	90.18
Precision	88.62	79.11	89.70
F-measure	87.97	84.83	89.94

Table 3: *Performance evaluation on STP dataset.*

Measure	BLSTM	PSF	Proposed
Recall	84.44	85.70	84.88
Precision	83.11	85.47	86.98
F-measure	83.74	85.61	85.92

Tables 1-3 give the performance of the proposed algorithm on TIMIT, NTIMIT and STP datasets compared against the state-of-the-art BLSTM [12] and perceptually significant features (PSF) [11] based methods. It can be observed from Table 1 that the proposed method performs equally well compared to BLSTM and PSF on TIMIT database. The speakers considered for evaluation on TIMIT dataset are different from those used for empirical analysis. Results given in Tables 2 and 3 show that the proposed method performs better than the other methods on channel degraded speech (NTIMIT) and conversational speech (STP dataset). It is important to note that the parameters (thresholds) of the proposed algorithm are kept constant for all the three datasets, whereas in case of BLSTM technique, separate models are trained for evaluation on TIMIT, NTIMIT and STP datasets. This shows that the proposed features capture the properties of vowel landmarks better.

Table 4 gives the performance of the proposed algorithm on TIMIT and NTIMIT datasets degraded with additive white noise. It can be observed from Table 4 that the proposed algorithm (for the same parameters) achieves nearly equal F-measure scores across different noise levels. This shows that the proposed features are also robust to additive noise. Analysis of STP in the presence of additive white noise is not performed as the dataset already contains different noises at varying noise levels.

Table 4: *Performance of proposed algorithm on TIMIT and NTIMIT added with white noise.*

	TIMIT			NTIMIT		
SNR	Recall	Precision	F-measure	Recall	Precision	F-measure
30 dB	90.94	94.64	92.75	90.06	89.65	89.85
20 dB	91.83	94.27	93.04	90.08	89.48	89.77
10 dB	93.37	90.01	91.66	93.63	83.11	88.06
5 dB	92.84	87.71	90.02	92.15	84.37	88.01
0 dB	89.49	87.70	88.59	86.33	84.91	85.61

6. Summary and conclusions

In this paper, acoustic features based on zero frequency filtered signal and envelope of the speech signal (obtained using single frequency filtering method) are proposed for robust detection of vowel landmarks in continuous speech. Using these features extracted around epoch locations, a rule-based algorithm is developed for vowel landmark detection. The performance of this algorithm is evaluated on three different datasets, i.e., TIMIT (read speech), NTIMIT (channel degraded speech) and Switchboard corpus (conversational speech). Evaluation results show that the proposed method performs equally well on read speech, and better on channel degraded and conversational speech, compared to the state-of-the-art methods. Also, the proposed method shows consistently good performance for different levels of noise degradations. This method doesn't require any prior training, and there is no need of parameter adjustment for new data.

Robust features to detect landmarks for other sound categories need to be defined in order to achieve landmark-based representation of speech signal, which forms our future work.

7. References

[1] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872-1891, Apr. 2002.

[2] A. W. Howitt, "Automatic syllable detection for vowel landmarks," Ph.D. thesis, Massachusetts Institute of Technology, USA, Jul. 2000.

[3] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," Ph.D. thesis, Massachusetts Institute of Technology, USA, 1994.

[4] A. Salomon, C. Y. Espy-Wilson and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1296-1305, Mar. 2004.

[5] IC Yoo and D. Yook, "Robust voice activity detection using the spectral peaks of vowel sounds," *ETRI journal*, vol. 31, no. 4, pp. 451-453, Aug. 2009.

[6] A. Juneja and C. Y. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1154-1168, Feb. 2008.

[7] A. Jansen, and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739-1758, Sep. 2008.

[8] M. Leena and B. Yegnanarayana. "Extraction and representation of prosodic features for language and speaker recognition," *Speech communication*, vol. 50, no. 10, pp. 782-796, Oct. 2008.

[9] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," *in Proc. Interspeech*, Pittsburgh, Pennsylvania, USA, Sep. 2006.

[10] Y. Zhang and J. Glass, "Speech rhythm guided syllable nuclei detection," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, pp. 3797-3800, Apr. 2009.

[11] A. A. Reddy, N. Chennupati and B. Yegnanarayana. "Syllable nuclei detection using perceptually significant features," *in Proc. Interspeech*, Lyon, France, pp. 963-967, Aug. 2013.

[12] C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg and B.Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, pp. 5256-5259, May 2011.

[13] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 8, pp. 2190-2201, Nov. 2007.

[14] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Washington, USA, pp. 945-948, May 1998.

[15] K. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[16] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, no. 3, pp. 267-281, May 2000.

[17] S. H. Dumpala, B. T. Nellore, R. R. Nevali, S. V. Gangashetty and B. Yegnanarayana, "Robust features for sonorant segmentation in continuous speech," *in proc. Interspeech*, Dresden, Germany, pp. 1987-1991, Sep. 2015.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus, Linguistic data consortium," Philadelphia, USA, 1993.

[19] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, New Mexico, USA, pp. 109-112, Apr. 1990.

[20] S. Greenberg, J. Hollenback and D. Ellis, "The Switchboard transcription project," *in LVCSR Summer Workshop Technical Reports*, Baltimore, Maryland, USA, vol. 72, pp. 104, 1996.

[21] Andrew Varga and Herman J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247-251, Jul. 1993.

[22] K. Murthy, B. Yegnanarayana and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469-472, Jun. 2009.

[23] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705-717, Feb. 2015.

[24] K. N. Stevens, "Acoustic phonetics," *MIT press*, Vol. 30, 2000.